Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jernej Porenta

Učenje iz besedilnih podatkovnih tokov za zaznavanje neželene elektronske pošte

magistrsko delo

magistrski študij informacijski sistemi in odločanje

Mentor: izr. prof. dr. Zoran Bosnić
Somentor: doc. dr. Mojca Ciglarič

Ljubljana, 2016

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Večna pot 113
1000 Ljubljana, Slovenija
telefon: 01 47 98 100
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št.: 133-MAG-ISO/2015
Datum: 08. 12. 2015

Jernej PORENTA, univ. dipl. inž. rač. in inf.

L j u b l j a n a

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Učenje iz besedilnih podatkovnih tokov za zaznavanje neželene elektronske pošte**

Learning from textual data streams for detecting email spam

Tematika naloge:

Sistemi za zaznavo neželenenih elektronskih sporočil uporabljajo različne metode za uvrščanje sporočil v kategoriji neželenih oziroma želenih elektronskih sporočil. Prva skupina teh metod analizira majhen nabor podatkov v zaglavju sporočila, druga pa celotno sporočilo z njegovo vsebino. Prve metode so manj natančne, vendar omogočajo prejemnikom, da z manjšo porabo računalniških virov določijo razred elektronskega sporočila (želena ali neželena pošta). Obenem te metode dovoljuje slovenska zakonodaja, ker ne posegajo v osebne podatke prejemnika oziroma naslovnika. Metode iz druge skupine so zahtevnejše za procesiranje in dosegajo manjše število napačno klasificiranih negativnih primerov ("false positives").

Kandidat naj v magistrski nalogi prouči možnosti za uvrščanje neželenih elektronskih sporočil na netradicionalen način – s prevedbo problema v inkrementalno učenje iz časovnih vrst. V delu naj iz elektronskih sporočil oblikuje atribute za strojno učenje, identificira najboljše in naj aplicira/evalvira različne metode za učenje iz podatkovnih tokov.

M e n t o r :

izr. prof. dr. Zoran Bosnić

S o m e n t o r i c a :

doc. dr. Mojca Ciglarič

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

D e k a n :

prof. dr. Nikolaj Zimic

Izjava o avtorstvu magistrskega dela

Spodaj podpisani Jernej Porenta sem avtor magistrskega dela z naslovom:

Učenje iz besedilnih podatkovnih tokov za zaznavanje neželene elektronske pošte (angl. *Learning from textual data streams for detecting email spam*)

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Zorana Bosnića in somentorstvom doc. dr. Mojce Ciglarič,
- so elektronska oblika magistrskega dela, naslov (slov., angl.), povzetek (slov., angl.) in ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela,
- soglašam z javno objavo elektronske oblike magistrskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 31. avgusta 2016 Podpis avtorja:

Najprej bi se rad zahvalil mentorju izr. prof. dr. Zoranu Bosniću in somentorici doc. dr. Mojci Ciglarič za odzivno, strokovno in nesebično pomoč pri izdelavi magistrskega dela. Zoranu gre dodatna zahvala zaradi izjemne vzpodbude in kolegialnega odnosa, ki je omogočil izdelavo tega magistrskega dela.

Za koristne nasvete, kolegialno pomoč in sproščeno delovno okolje se želim zahvaliti kolektivu Akademske in raziskovalne mreže Slovenije, ki mi je omogočilo raziskovanje področja neželene elektronske pošte.

Izredna zahvala gre gdč. Zdenki Velikonja, ki mi je vedno pomagala s koristnimi nasveti okrog vseh postopkov in prenašala moje odzive na to.

Rad bi se zahvalil tudi vse prijateljem za moralno podporo. Barbara, brez tebe mi ne bi uspelo.

Za konec bi se rad zahvalil svojim staršem, bratu in sestri, ki so mi vse to omogočili.

Mojim staršem.

Kazalo

Slike

Tabele

Seznam uporabljenih kratic

kratica	angleško	slovensko
CA	classification accuracy	klasifikacijska točnost
SMTP	simple mail transport protocol	protokol za izmenjavo elektronskih sporočil
DKIM	domain-keys identified email	metoda domenskega podpisovanja elektronske pošte
SPF	sender policy framework	metoda preverjanja pošiljatelja elektronske pošte
VFDT	very fast decision trees	algoritem Hoeffdingovih dreves
RBL	real/time blackhole list	seznam slabih IP naslovov
DNS	domain name system	sistem domenskih imen
DNSBL	DNS block list	seznam slabih IP naslovov v sistemu DNS
DMARC	domain message authentication reporting and conformance	sistem podpisovanja in preverjanja domenskih podpisov elektronske pošte
IP	internet protocol	internetni protokol
TCP	transmission control protocol	prenosni kontrolni protokol
CSV	comma separated values	z vejico ločene vrednosti

ARNES	academic and research network of Slovenia	Akademska in raziskovalna mreža Slovenije
ASN	autonomous system number	samostojna številka sistema
SLING	Slovenian grid initiative	slovenska iniciativa za nacionalni grid
FQrDNS	forward-confirmed reverse DNS	potrjen obratni DNS zapis
MX	DNS mail exchange	DNS zapis poštnega strežnika

Povzetek

Naslov: Učenje iz besedilnih podatkovnih tokov za zaznavanje neželene elektronske pošte

V magistrski nalogi je predstavljena metoda uvrščanja sporočil v kategoriji neželenih oziroma želenih elektronskih sporočil s prevedbo problema v inkrementalno učenje iz časovnih vrst. Razširjeni sistemi za uvrščanje neželene elektronske pošte uporabljajo predvsem metode paketnega učenja (naivni Bayesov klasifikator), medtem ko je v magistrski nalogi predstavljeno uvrščanje z uporabo metod analize tokov.

Za učenje smo tako izbrali atribute, ki ne vsebujejo osebnih podatkov in za katere ni treba pridobiti privoljenja pošiljatelja oziroma prejemnika (atributi sestavljeni iz ovojnice elektronske pošte). S pomočjo algoritmov za učenje iz podatkovnih tokov (VFDT, cVFDT) smo zaporedje elektronskih sporočil obravnavali kot besedilni tok podatkov. Rezultate smo primerjali s tradicionalnimi metodami označevanja neželene elektronske pošte in ugotovili, da metode inkrementalnega učenja iz podatkovnih tokov na primeru problemske domene uvrščanja neželene elektronske pošte dosegajo manjšo klasifikacijsko točnost in so zato manj primerne za uporabo.

Ključne besede: elektronska pošta, strojno učenje, analiza podatkovnih tokov.

Abstract

Title: Learning from textual data streams for detecting email spam

This master thesis introduces a method for the detecting email spam through the translation problem in incremental learning of the time series. Common spam detection systems mainly use methods of supervised learning (naive Bayesian classifier, decision trees), while in the master's thesis presents the classification by using the methods of data stream mining.

For learning sets, we also choose the attributes that do not contain personal data and which are not required to obtain the consent of the sender or the recipient (attributes consist the envelope part of e-mail). With the help of algorithms for learning from data streams (VFDT, cVFDT) we used the electronic sequence of messages as text data stream. The results were compared with the traditional spam detection methods and they show that traditional spam detection methods have higher accuracy compared to algorithms for learning from data stream and therefore are not suitable for detecting email spam.

Keywords: email, machine learning, stream mining.

Poglavje 1

Uvod

Elektronska pošta je dandanes eden vodilnih načinov elektronske komunikacije. Trenutno število uporabnikov elektronske pošte presega 2.6 miljarde in pričakovati je, da bo do konca leta 2019 vsaj ena tretjina svetovnega prebivalstva uporabljala elektronsko pošto [1]. Poleg velikega števila uporabnikov elektronske pošte je pomembno tudi to, da količina elektronske pošte še vedno narašča in dnevno že presega 200 miljard poslanih elektronskih sporočil.

S takšno količino dnevno poslanih sporočil tovrstna komunikacija predstavlja zanimivo področje za oglaševanje in zaradi različnih lastnosti elektronske pošte tudi nelegalne aktivnosti. Prav zaradi teh lastnosti delimo elektronsko pošto v dve kategoriji:

1. želeno elektronsko pošto (angl. ham): to so elektronska pošta z vsebino, ki jo je prejemnik želel (osebna sporočila, poslovna sporočila) oziroma se je nanjo naročil (distribucijski seznami, oglasi ...);

2. neželeno elektronsko pošto (angl. spam): to so elektronska sporočila, navadno z oglaševalsko vsebino, ki je prejemnik ne želi dobivati in pošiljatelj nima dovoljenja prejemnika za pošiljanje te elektronske pošte.

Problem neželene elektronske pošte je v tem, da predstavlja več kot 53% vseh poslanih sporočil [2]. Neželena elektronska pošta tako utaplja koristno komunikacijo, obenem pa večino stroškov nosi sam prejemnik. Pogosto upo-

rabniki elektronske pošte med neželena sporočila uvrščajo tudi posredovana sporočila, samodejna sporočila in obvestila o napakah, vendar ta ne spadajo mednje.

Najbolj tipični primeri neželene elektronske pošte so:

1. oglaševanje in prodaja izdelkov, ki nas ne zanimajo: prodaja zdravil s črnega trga, prodaja ponaredkov, oglaševanje brez naše privolitve ...;
2. virusi in zlonamerna koda: elektronska sporočila, ki vsebujejo programe, ki lahko škodujejo delovanju računalnika;
3. spletne prevare: elektronska sporočila, ki s prevaro pošiljatelju omogočijo finančno ali materialno korist v škodo prejemnika.

Neželena elektronska pošta je zanimiva za nelegalne aktivnosti predvsem zaradi pomanjkljivosti celotnega sistema elektronske pošte in protokola SMTP [3]. Te so predvsem:

- ponarejanje pošiljateljevega imena ali elektronskega naslova: protokol SMTP omogoča ponarejanje pošiljateljevega imena ali elektronskega naslova, kar pošiljatelji neželene elektronske pošte s pridom izkoriščajo;
- prikrivanje vira: poleg ponarejanja pošiljatelja lahko s protokolom SMTP priredimo dejansko pot do vira elektronske pošte in s tem zmedemo prejemnika;
- poceni oglaševanje: cena elektronske pošte je izjemno nizka, saj vse stroške plača prejemnik (prenos podatkov, elektronski naslov ...);
- ciljanje na zasebnost prejemnika: velikokrat neželena pošta naslavlja teme, ki jih ljudje prepoznamo za bolj zasebne: težave v spolnosti, telesna teža, luksuzni izdelki ...

Sami viri neželene elektronske pošte so največkrat omrežja zlorabljenih strežnikov – angl. botneti. V letu 2010 so taka omrežja zlorabljenih strežnikov poslala 88% vse neželene elektronske pošte [4, 5] in še danes predstavljajo

večinski delež virov neželene elektronske pošte. Poleg omrežij zlorabljenih strežnikov med večje pošiljatelje prištevamo še ukradena uporabniška imena veljavnih pošiljateljev in tudi nevešče oglaševalce, ki ne upoštevajo zakonskih omejitev glede oglaševanja in pridobivanja naslovnikov za njihove oglaševalke kampanje.

Sistemi za zaznavo neželenenih elektronskih sporočil uporabljajo različne metode za uvrščanje sporočil v kategoriji neželenih oziroma želenih elektronskih sporočil [6], ki jih delimo v dve skupini:

1. metode, ki uvrščajo sporočila pred samim prejemom, kjer je količina podatkov o sporočilu precej majhna (uporaba seznamov veljavnih in neveljavnih elektronskih naslovov pošiljateljev; uporaba seznamov naslovov IP znanih pošiljateljev; uporaba mehanizmov ugleda IP) [7, 8, 9];

2. metode, ki uvrščajo sporočila po prejemu celotnega sporočila (analiza zaglavja in vsebine elektronskega sporočila; uporaba avtentikacije pri pošiljateljih elektronskega sporočila; elektronski podpisi DKIM, SPF; uporaba mehanizmov izziva in odgovora – (angl. "challenge-response"); uporaba porazdeljenih mehanizmov kontrolnih vsot; uporaba algoritmov umetne inteligence) [10, 11, 12, 13, 14, 15].

Metode, ki spadajo v prvo skupino, so manj natančne [12], vendar omogočajo prejemnikom, da z manjšo porabo računalniških virov določijo razred (kategorijo) elektronskega sporočila. Obenem jih dovoljuje slovenska zakonodaja, ker ne posegajo v osebne podatke prejemnika oziroma naslovnika.

Metode iz druge skupine so zahtevnejše za procesiranje, a so istočasno zaradi analize same vsebine bolj zanesljive in s tem dosegajo manjše število napačno klasificiranih negativnih primerov ("false positives"). Ti so v zaznavi neželenih elektronskih sporočil najbolj razširjena merila uspešnosti sistema za zaznavo neželene elektronske pošte.

Velik problem samodejnega prilagajanja obliki neželenih sporočil obenem ostaja problem časovnega okvira, v katerem se ta elektronska sporočila po-

javljajo. V praksi je namreč veliko primerov, kjer pošiljatelji neželenih elektronskih sporočil izkoristijo obliko obstoječih veljavnih elektronskih sporočil in v njo podtaknejo svojo vsebino [16, 17]. Sistemi za zaznavo neželenih elektronskih sporočil morajo biti zato pripravljeni na včasih hitre, včasih pa počasne spremembe pri prejemanju neželenih elektronskih sporočil [18, 19]. Samodejno prilagajanje zaznavanja neželenih elektronskih sporočil ima manjšo klasifikacijsko točnost tudi v primerih, ko je vzorec neželenih elektronskih sporočil majhen in so le-ta namenjena zgolj manjšemu, natančno izbranem vzorcu elektronskih naslovov [20, 21, 22].

Zaradi vseh zgoraj omenjenih težav pri pravilnem razvrščanju elektronske pošte smo v tej magistrski nalogi uporabili nov pristop, ki upošteva elektronsko pošto (in njene atribute) kot tok podatkov v časovni vrsti, iz katerega se s pomočjo znanih algoritmov za strojno učenje iz podatkovnih tokov učimo, da lahko natančneje obidemo omejitve znanih metod.

Poleg samega pravilnega razvrščanja elektronske pošte smo velik poudarek namenili atributom, ki so na voljo v procesu učenja. Slovenska zakonodaja namreč ponudnika elektronske pošte omejuje, da sme pregledovati vsebino samo tiste elektronske pošte za katero je naslovnik podelil izrecno soglasje. Če tega soglasja ponudnik ni pridobil, sme uporabiti samo podatke iz ovojnice elektronske pošte (po protokolu SMTP) in s tem se klasifikacijska točnost zmanjša. Prav zaradi teh omejitev smo se v magistrski nalogi odločili upoštevati samo tiste atribute, za katere ponudnik elektronske pošte ne potrebuje soglasja uporabnika.

Magistrska naloga je razdeljena v več poglavij. V uvodu je predstavljena problemska domena uvrščanja neželene elektronske pošte, v drugem poglavju so predstavljene metode analize podatkovnih tokov z uporabo algoritmov za inkrementalno učenje. Pregled področja in že opravljenih pristopov je predstavljen v poglavju tri. Postopek izgradnje podatkovne zbirke smo predstavili v naslednjem poglavju, medtem ko smo v poglavju pet analizirali podatkovno zbirko, primerjali algoritme za strojno učenje in jim izbrali optimalne parametre. Povzetek ugotovitev je predstavljem v zaključku naloge. V dodatek

magistrske naloge smo dodali še natančneje predstavljene rezultate iz poglavja analize.

.

Poglavje 2

Pregled področja

2.1 Pregled metod za uvrščanje elektronske pošte

V literaturi je problem označevanja neželene elektronske pošte precej raziskano področje. Od razmaha elektronske pošte z uveljavitvijo Interneta v devetdesetih letih 20. stoletja poznamo kar nekaj prebojnih dogodkov pri označevanju neželene elektronske pošte.

Prvo razvrščanje elektronske pošte je bilo na ramenih navadnih prejemnikov, ki pa so kaj kmalu ugotovili, da jim to vzame preveč časa (in posledično denarja, [4]) in zato so razvili prve sisteme za označevanje in razvrščanje neželene elektronske pošte. Prvi sistemi za samodejno razvrščanje neželene elektronske pošte so uporabljali primitivne metode, ki so temeljile prevsem na enostavnih pravilih, ki so jih sestavili sami uporabniki elektronske pošte. O kakovosti teh pravil v literaturi ni podatkov, saj je to metodo težko izmeriti kot tudi primerjati z drugimi. Iz teh sistemov so izšli sistemi, ki uvrščajo elektronsko pošto na podlagi vnaprej pripravljenih pravil. Najbolj pogosto uporabljeno orodje, ki je še danes temelj odprtokodnega razvrščanja neželene elektronske pošte, je SpamAssassin [23]. Vsako pravilo, ki ga SpamAssassin vsebuje, je ovrednoteno z določeno utežjo in vsota vseh pravil, ki se ujamejo na posameznem elektronskem sporočilu, pravilno razvrsti elektronsko sporo-

čilo. Problem uporabe teh javno dostopnih pravil je v tem, da jih poznajo tudi pošiljatelji in se jim poskušajo izogniti.

Nov pristop v razvrščanju elektronske pošte je bila tudi vpeljava črnih in belih seznamov pošiljateljev [24], kjer na podlagi zgodovine pošiljanja ustvarimo sezname veljavnih in neveljavnih pošiljateljev. Z uporabo dodatnih metod umetne inteligence lahko take sezname precej izpopolnimo in zato je dandanes možno uporabiti zelo natančne sezname pošiljateljev neželene elektronske pošte. Ti seznami uporabljajo sistem domenskih imen in na tak način omogočajo enostavno vključitev v obstoječe sisteme za označevanje neželene elektronske pošte. Eden prvih takih seznamov je bil *Real-time Blackhole List (RBL) for Mail Abuse Prevention System (MAPS)* [25]. Dandanes jih poznamo veliko, ki se po kakovosti precej razlikujejo [26].

Nadgradnja seznamov je bila tudi uporaba skupnih baz znanja o elektronski pošti. Običajno v bazah znanja dobimo *podpise* zgoščenih vrednosti sporočil. Pošiljatelji neželene elektronske pošte namreč razpošljejo velike količine enakih elektronskih sporočil, kar pomeni, da lahko nekateri naslovniki že predhodno označijo ta sporočila za neželena in ostali prejemniki uporabijo to informacijo ob prejemu [27, 28]. Problem takega sistema je v tem, da je njegova točnost odvisna od števila pošiljateljev in tudi od količine poslane neželene elektronske pošte. Včasih so pošiljatelji poslali velike količine take pošte, medtem ko se danes (tudi z namenom izogibanja) raje osredotočijo na manjše skupine naslovnikov in tako obidejo, da bi se informacija o njihovih sporočilih znaša v bazah znanja o elektronski pošti. Obenem je problem teh baz tudi to, da pošiljatelji personalizirajo poslana sporočila in s tem begajo baze. V izogib temu je treba skrbno izbrati atribute posameznega sporočila, da ga lahko objavimo v javno dostopni bazi o elektronski pošti [29, 30].

Nov pristop je bila tudi uvedba metode izziv-odgovor (angl. "*challenge-response*"), kjer mora pošiljatelj izvesti vnaprej predvidene korake preden bo pošiljanje uspešno. Ti koraki so lahko enostavni, a obenem dovolj računsko zahtevni, da jih pošiljateji velike količine elektronske pošte nimajo časa reševati [31, 32, 33]. Osnovni princip te metode, ki je največkrat v uporabi,

je uporaba avtentikacije in avtorizacije pred samim pošiljanjem elektronske pošte.

Med precej uspešne metode preprečevanja dostave neželene elektronske pošte prištevamo tudi metodo zaustavljanja ("greylisting", [34]). Ta metoda deluje na principu strogega upoštevanja protokola SMTP, ki predvideva večkratno pošiljanje v primeru, da je bilo sporočilo prvič zavrnjeno. Namen pošiljateljev neželene elektronske pošte je namreč, da v kratkem časovnem obdobju pošljejo čim večjo količino neželene elektronske pošte. Z uporabo metode zaustavljanja pošiljatelju ustavimo prvo pošiljanje s sporočilom o napaki pri dostavi. Veljavni pošiljatelji tako napako upoštevajo in ponovno poskušajo dostaviti sporočilo, medtem ko pošiljatelji neželene pošte običajno nimajo zmogljivosti, da bi lahko upoštevali take napake in tako sporočilo posledično ni dostavljeno.

V zadnjem času so na plan prišle tudi bolj tehnične metode, ki razširjajo sam protokol SMTP z uporabo dodatnih mehanizmov, ki bolje identificirajo pošiljatelje želene elektronske pošte. Med te metode prištevamo:

- Sender policy framework – SPF [35], ki s pomočjo uporabe sistema DNS prejemniku omogoča enostavno preverjanje pošiljatelja in elektronske pošte, ki jo pošilja;
- Sender-ID [36], ki razširja delovanje sistema SPF z uporabo dodatnih podatkov,
- DomainKeys [37] uporablja metodo uporabe infrastrukture javnih ključev skozi sistem DNS. Vsako elektronsko sporočilo se z uporabo te metode podpiše in prejemnik lahko ob prejemu preveri avtentičnost sporočila,
- DomainKeys Identified Mail – DKIM [11] združuje metodi Sender-ID in DomainKeys in je danes najbolje uporabljena metoda za preverjanje avtentičnosti pošiljatelja. Metoda v zaglavje elektronske pošte vstavi podpis, ki ga lahko prejemnik preveri ob prejetju z uporabo infrastrukture javnih ključev,

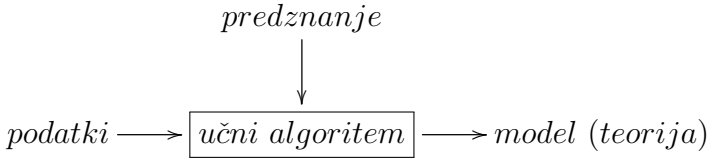
- Domain-based message authentication, reporting and conformance – DMARC [38] je zadnja metoda, ki združuje vse prej omenjene metode in jih razširja z možnostjo povratne informacije, ki lahko pošiljateljem pomaga pri boju z neželeno elektronsko pošto.

2.2 Uporaba metod strojnega učenja pri uvrščanju elektronske pošte

Strojno učenje je področje umetne inteligence in ga uporabljamo predvsem za analizo podatkov, odkrivanje zakonitosti v podatkih, gradnjo napovednih modelov. Metode strojnega učenja lahko uporabljamo za reševanje različnih problemov: klasifikacije, regresijo, povezovalna pravila in logične relacije, odkrivanje zakonitosti, nenadzorovano učenje in spodbujevalno učenje. Pri nadzorovanem strojnem učenju je podana končna podatkovna množica na podlagi katere lahko izdelamo odločitveni model, ki ga lahko uporabimo na neznanih podatkih.

Podatki za strojno učenje so pridobljeni iz različnih virov in zaradi velike količine, oblike ali uporabnikovega vpliva predstavljajo izzive pri samem strojnem učenju.

Na področju elektronske pošte se največkrat uporablja metode uvrščanja z uporabo različnih, že znanih metod: odločitvenih dreves, naivnega Bayesovega klasifikatorja, klasifikatorja z najbližjimi sosedi, nevronskimi mrežami in hibridnimi metodami. Strojno učenje se tako v elektronski pošti lahko uporablja v vseh fazah pošiljanja le-te [39] in z njim lahko uspešno pravilno



Slika 2.1: Algoritem za strojno učenje

razvrstimo nad 90% vse elektronske pošte [40].

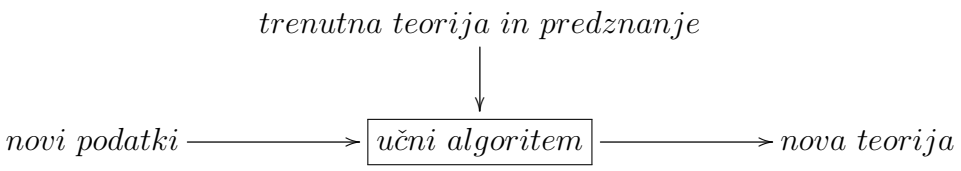
Sama elektronska pošta zaradi svoje specifike povzroča strojnemu učenju precej težav [41] zaradi:

- različne porazdelitve obeh razredov: neželena elektronska pošta še vedno predstavlja večino poslane pošte [2],
- različne cene napačne klasifikacije: napačno razvrščanje v enega izmed razredov je različno pomembno,
- pogostega prilagajanja pošiljateljev: pošiljatelji pogosto spreminjajo samo vsebino elektronske pošte in načine pošiljanja in s tem znižujejo klasifikacijsko točnost algoritmov strojnega učenja.

Danes imamo za vse te omejitve različne rešitve: prilagojen način učenja algoritmov z uporabo delno usmerjenega učenja [42], ki predvideva da algoritme najprej naučimo z manjšim naborom dobro označenih elektronskih sporočil obeh razredov, ki jih kasneje razširjamo z dodatnimi primeri. Zaradi spreminjanja oblike in vsebine elektronskih sporočil Cruz in Cormack [44] izpostavita tudi problem učenja iz starih primerov elektronske pošte, ki lahko povzroči precej nizko klasifikacijsko točnost. Problem različne cene napačne klasifikacije so deloma razrešili Androutsopoulos, Magirou in Vassilakis [45], boljšo rešitev so predlagali tudi Chan, Koprinska in Poon [42], ki so upoštevali tudi končnega prejemnika elektronske pošte in s tem povezali, da ljudje različno ocenjujemo ceno napačne klasifikacije.

Algoritmi za paketno učenje iz stacionarnih podatkov (slika 2.1) predvidevajo, da je količina podatkov končna in da jo algoritem lahko prebere večkrat in s tem izboljša trenutni odločitveni model. Podatki se pri tem ne spreminjajo, algoritmi jih lahko preberejo večkrat zaporedoma.

Na drugi strani inkrementalno učenje iz podatkovnih tokov uporabi podatke kot neskončni podatkovni tok, ki ga lahko beremo samo zaporedno in le enkrat. Zato imamo pri inkrementalnem učenju iz podatkovnih tokov določene omejitve:



Slika 2.2: Inkrementalno učenje

- količina podatkov je neskončna in samo del teh podatkov lahko shranimo, medtem ko ostale zavržemo,
- podatki morajo biti obdelani v najkrajšem možnem času in so po obdelavi zavrženi,
- podatki se s časoma spreminjajo in zato starejši podatki postajajo nepomembni (včasih celo škodljivi) za trenutni model.

Zato pri strojnem učenju iz podatkovnih tokov uporabimo algoritme za inkremenalno učenje (slika 2.2), kjer se model spreminja ob vsakem novem prejetem učnem primeru. Odločitveni model se tako prilagaja novim učnim primerom, medtem ko lahko starim učnim primerov zmanjšuje vlogo. S tem se lahko klasifikacijska točnost ohranja ali tudi izboljšuje.

Na temo klasifikacijskih problemov v področju analize podatkovnih tokov je bilo uporabljenih že kar nekaj inkrementalnih odločitvenih dreves, ki upoštevajo omejitve podatkovih tokov. Tako sta Domingos in Hulten preučevala odločitvena drevesa na podatkovnih tokovih in za to razvili algoritem VFDT [46], ki temelji na sistemu Hoeffdingovih dreves.

Problem pri inkrementalnem učenju iz podatkovnih tokov predstavljajo spremembe porazdelitve v podatkovnem toku. Te spremembe lahko povzročijo padec klasifikacijske točnosti.

V primeru elektronske pošte so take spremembe pogostejše, saj pošiljatelji neželene elektronske pošte nenehno spreminjajo obliko in načine pošiljanja, da s tem povzročijo napačno klasifikacijo elektronskega sporočila. Problem prilagajanja sprememb v podatkovnem toku v algoritmu VFDT so Hulten,

Spencer in Domingos izboljšali z algoritmom cVFDT [47], ki se bolje prilagaja zveznim atributom in s tem poveča klasifikacijsko točnost. Algoritem cVFDT gradi vzporedno drevo VFDT algoritmu in takoj, ko obstoječe drevo ne dosega več dobre klasifikacijske točnosti, ga zamenja z vzporednim. Algoritma VFDT in cVFDT sta natančneje predstavljena v poglavju 2.2.1. Celoten sistem za prilagajanje sprememb v podatkovnem toku je predlagal tudi Wang in sod., ki je z uporabo uteži rezultatov različnih algoritmom (C4.5, naivni Bayes, RIPPER) [48] nad učnimi podatki dosegel boljše prilagajanje spremembam in s tem izboljšano klasifikacijsko točnost.

Problem uvrščanja elektronske pošte so poskušali z uporabo inkrementalnega učenja iz podatkovnih tokov rešiti Carmona-Cejudo in sod. [49, 13]. Elektronska sporočila s celotno vsebino so predstavili kot tok podatkov in s pomočjo različnih algoritmov za inkrementalno učenje preverili klasifikacijsko točnost. Pristop k problemu je precej podoben našemu pristopu, vendar se razlikuje v lastnostih podatkovnega toka. Medtem ko je naš pristop usmerjen k ponudnikom elektronske pošte, ki brez pridobljenega soglasja prejemnika ne smejo pregledovati, je pristop Carmona-Cejudo in sod. usmerjen predvsem h končnemu uporabniku in v tem oziru se razlikujejo lastnosti podatkovnega toka. Podatkovni tok, ki ga lahko uporabi ponudnik elektronske pošte, namreč ne sme vsebovati nobenih osebnih podatkov in same vsebine elektronskega sporočila. Na drugi strani lahko končni uporabnik uporabi vse podatke iz elektronskega sporočila, vključno z vsebino.

Zato predlagan pristop uporablja samo tiste atribute iz podatkovnega toka, ki ne vsebujejo osebnih podatkov prejemnika in pošiljatelja. Razlika v našem pristopu je tudi v uporabi dodatnih atributov, ki so na voljo ponudnikom elektronske pošte in ne posegajo v zasebnost prejemnika. Teh informacij pristop Carmona-Cejudo in sod. niso uporabili.

2.2.1 Pregled algoritmov VFDT in cVFDT

Algoritma VFDT in CVFDT sta med najbolj pogosto uporabljenimi algoritmi za učenje inkrementalnih odločitvenih dreves iz podatkovnih tokov.

Razvila sta ju Domingos in Hulten [46] in temeljita na Hoeffdingovih drevesih, ki uporabljajo Hoeffdingovo mejo kot kriterij za delitev lista. Le-ta je definirana s formulo 2.1:

$$\varepsilon = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2n}} \tag{2.1}$$

Algoritem ob vsakem novem primeru oceni kakovost atributov in jih rangira od najbolj do najmanj pomembnega (A1, A2, ...). Če je razlika kriterijskih funkcij (npr. informacijski prispevek) atributov A1 in A2 večja od ε, lahko z verjetnostjo $1 - \delta$ trdimo, da je atribut A_i primeren za deljenje.

Algoritem VFDT ne uporablja nobene metode za zaznavo sprememb porazdelitev v podatkovnem toku in zato predvideva, da se porazdelitve v podatkovnem toku ne spreminjajo. Če so take spremembe pogoste, dosega algoritem VFDT zato nižjo klasifikacijsko točnost.

Problem zaznave sprememb porazdelitev v podatkovnem toku algoritma VFDT rešuje algoritem CVFDT, ki uporablja drseče okno za oceno točnosti vej v odločitvenem drevesu. Če klasifikacijska točnost vzporedno zgrajenega drevesa iz podatkov drsečem oknu preseže tisto v odločitvenem drevesu, algoritem zamenja tiste dele odločitvenega drevesa, ki imajo nižjo klasifikacijsko točnost. Na tak način se algoritem CVFDT bolje prilagaja podatkovnim tokovom, v katerem so pogoste spremembe porazdelitve.

Poglavje 3

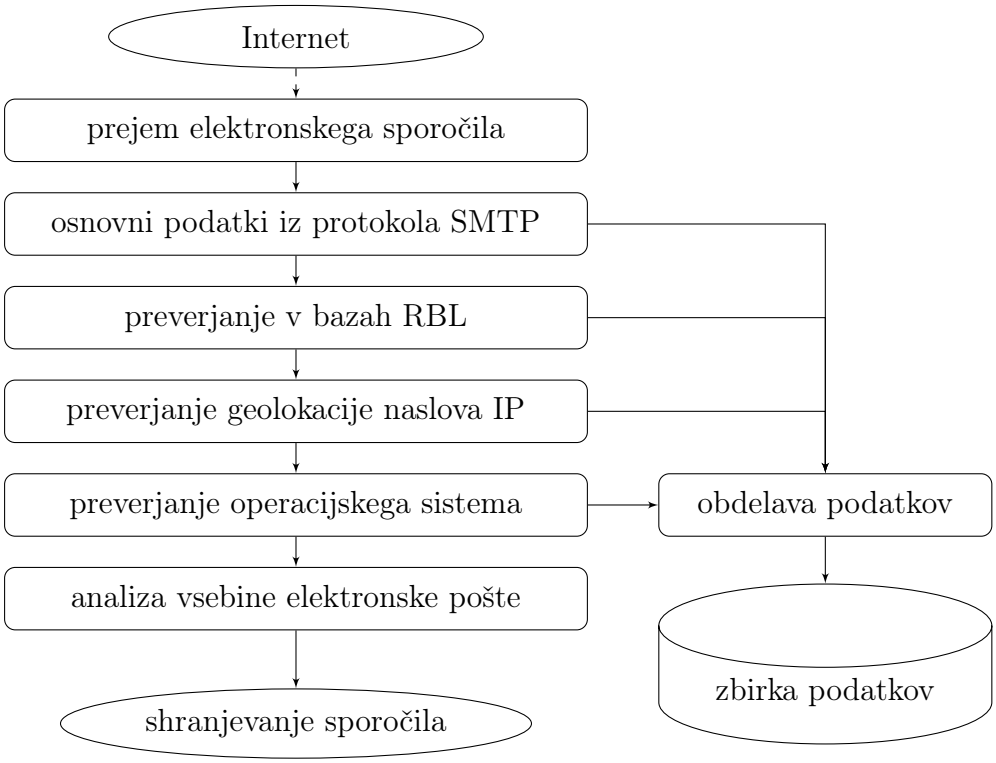
Priprava podatkovnega toka

Za potrebe analize je bilo treba pridobiti podatke o elektronski pošti, jih primerno urediti v zbirno obliko, izbrati najbolj primerne atribute in jih nato obdelati z algoritmi za učenje iz podatkovnih tokov.

Izhodišče za zbiranje podatkov je temeljilo na tem, da podatki ne smejo posegati v pravice pošiljateljev ali prejemnikov in s tem pri zbiranju ni treba pridobiti neposrednega soglasja pošiljatelja ali prejemnika elektronske pošte. V praksi to pomeni, da smemo v zbirko podatkov uvrstiti:

- podatke, ki so bili zbrani ob prenosu elektronske pošte: čas prejema, velikost sporočila, naslov IP pošiljatelja/prejemnika, operacijski sistem pošiljatelja, hitrost povezave in drugo,
- podatke iz ovojnice SMTP: pošiljatelj, prejemnik, ime pošiljateljevega strežnika in drugo,
- pridobljene podatke iz javnih zbirk: podatki DNSBL, geolokacija pošiljatelja in drugo,
- podatke, izračunane iz ostalih podatkov: časovni pas pošiljatelja, ujemanje med reverzno preslikavo in naslovom IP in drugo.

Med zbrane podatke pa ne smemo uvrstiti vsebine elektronske pošte, kar pomeni, da ne moremo preveriti elektronskega podpisa DKIM, izvesti analize vsebine in s tem se število podatkov občutno zmanjša.



Slika 3.1: Sistem za zbiranje podatkov

Glede na omejitve smo zasnovali sistem, ki je lahko zbral vse podatke, ne da bi posegal v pravice pošiljateljev in prejemnikov. Ta je predstavljen na sliki 3.1. Podatki, zbrani na ta način ne vsebujejo osebnih podatkov prejemnikov in tako pri zbiranju ne posegamo v njihove pravice. Tako dobimo ločeno zbirko podatkov, ki ne vsebuje osebnih podatkov, medtem ko se sama sporočila shranijo ločeno v predale prejemnikov elektronske pošte. Celoten sistem za zbiranje podatkov temelji na odprtokodnih rešitvah, ki so bile prilagojene posebej za ta sistem:

- prejem elektronskega sporočila: programska oprema za strežnik SMTP postfix[1];
- preverjanje v bazah DNSBL in preverjanje geolokacije naslova IP: programska razširitev za postfix: ppolicy[2];

[1] `http://www.postfix.org`

[2] `http://kmlinux.fjfi.cvut.cz/~vokacpet/activities/ppolicy/`

- preverjanje operacijskega sistema: programska oprema za zaznavo operacijskega sistema iz podatkov povezave TCP – p0f[3]

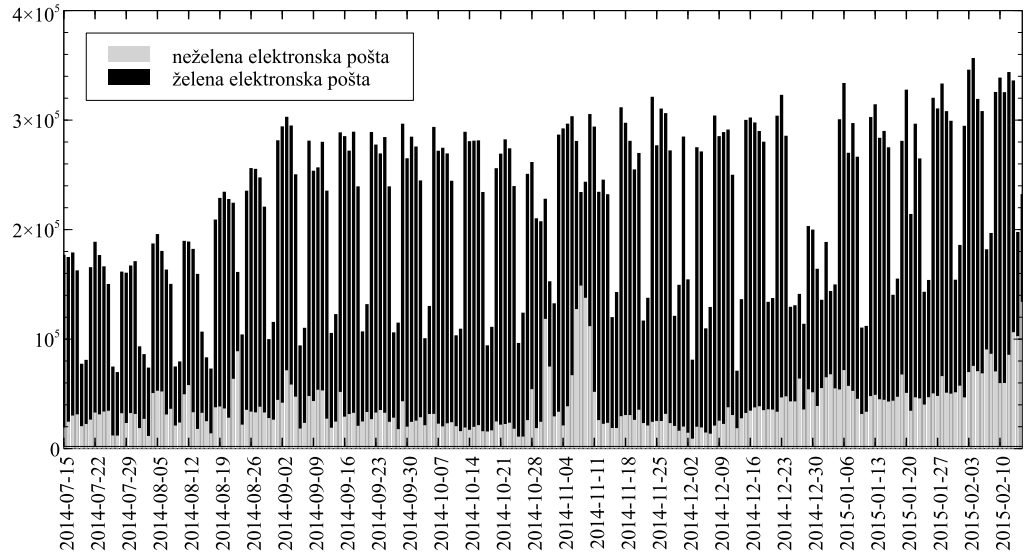
Pri obdelavi podatkov o elektronski pošti je zelo pomembna časovna komponenta obdelave, saj se mnogi podatki iz javno dostopnih baz (RBL ...) precej hitro spreminjajo in s tem spreminjajo svojo informacijsko vrednost pri zaznavi neželene elektronske pošte. Velikokrat se namreč zgodi, da javne dostopne baze še nimajo podatka o tem, da bi nek naslov IP pošiljal neželeno elektronsko pošto, medtem ko je ta naslov IP že vir take pošte. Predvsem zaradi tega razloga se vsi podatki obdelajo in shranijo ob prejemu elektronske pošte, tako da na tak način vsebujejo vse informacije, ki so bile takrat na voljo. S pomočjo celotnega sistema smo zbrali podatke o vsaki prejeti elektronski pošti in vse pripadajoče atribute. Dodatno smo te atribute razširili s podatkom o vsebini elektronske pošte, ki smo ga dobili kot povratno informacijo iz sistema za označevanje neželene elektronske pošte. Ta lahko preveri vsebino elektronske pošte z dovoljenjem prejemnika in s tem precej izboljša klasifikacijsko točnost. Uporabljen sistem za osnovo vzame programsko opremo amavisd-new[4], ki predstavlja ogrodje za zaznavo neželene elektronske pošte. V to ogrodje je vpeta programska oprema SpamAssassin [23], CRM114[5], dspam[6], ClamAV[7] kot tudi dodatna pravila, ki smo jih s pomočjo avtorja programske opreme amavisd-new[8], razvili za povečanje klasifikacijske točnosti. S tako sestavljenim sistemom smo dosegli klasifikacijsko točnost višjo od 99% [7].

Sistem za zbiranje podatkov je tako zbiral podatke v obdobju med 15. julijem 2014 in 15. februarjem 2015 za elektronske naslove uporabnikov storitev elektronske pošte omrežja ARNES[9] in jih shranjeval v obliko CSV.

Po končanem zbiranju podatkov smo podatke obdelali v obliko, primerno

[3] `http://lcamtuf.coredump.cx/p0f3/`
[4] `https://www.ijs.si/software/amavisd/`
[5] `http://crm114.sourceforge.net/`
[6] `http://dspam.nuclearelephant.com/`
[7] `https://www.clamav.net/`
[8] Mark Martinec - mark.martinec@ijs.si
[9] ARNES – Akademska in raziskovalna mreža Slovenije, `http://www.arnes.si/`

za nadaljno obravnavo. To pomeni, da smo podatke časovno uredili, uredili v kodno tabelo UTF8 in vse atribute uredili v skladu z opisano zalogo vrednosti v tabeli 6.1. Tako smo dobili podatkovno zbirko, ki je predstavljena na sliki 3.2



Slika 3.2: Število želenih in neželenih elektronskih sporočil v obdobju med 15. julijem 2014 in 15. februarjem 2016

Popolna podatkovna zbirka je vsebovala 47.073.461 zapisov o prejetih elektronski sporočilih od tega 8.475.499 neželenih in zasedla skoraj 27 gigabajtov podatkov.

Vsak dan je tako v povprečju sistem prejel 217.932 elektronskih sporočil (s standardnim odklonom 78.714) od katerih jih je bilo v povprečju 39.238 neželenih.

Poglavje 4

Analiza podatkov

Po končanem zbiranju podatkov smo izvedli analizo podatkov in pri tem uporabili metode strojnega učenja. Uporabili smo dva različna pristopa: metode za paketno učenje iz stacionarnih podatkov in zatem metode učenja iz podatkovnih tokov.

Postopek analize podatkov smo ločili v štiri ločene korake:

- izbor 20 najbolj pomembnih atributov učne množice;
- izbor in primerjava algoritmov za učenje iz stacionarnih podatkov in algoritmov za učenje iz podatkovnih tokov;
- izbor velikosti učnega okna pri metodah strojnega učenja iz podatkovnih tokov;
- primerjava z obstoječimi metodami za označevanje neželene elektronske pošte.

4.1 Izbor najbolj pomembnih atributov

Zaradi velike podatkovne zbirke s 83 atributi smo v prvem koraku omejili število atributov in s tem izboljšali klasifikacijsko točnost kasneje uporabljenih metod strojnega učenja in obenem pridobili domensko znanje o pomembnih atributih.

Za oceno kvalitete atributov v dvorazrednih klasifikacijskih problemih smo uporabili funkcijo ReliefFexpRank [50], ki učinkovito rešuje problem odvisnosti atributov, ki bi se lahko pojavili v naši podatkovni zbirki. Pri tem smo uporabili funkcijo ReliefFexpRank iz programskega paketa COREl earn [51] v okolju R.

Zaradi velike zaloge vrednosti posameznih atributov smo pred začetkom izbora 20 najbolj pomembnih atributov uporabili zgoščevalno funkcijo, ki je atribute z veliko zalogo vrednosti uvrstila v diskretne atribute z zalogo vrednosti 1024 različnih vrednosti. Ti atributi so: *ASN, SMTP HELO ime, ime pošiljatelja* in *naslov IP pošiljatelja*.

Zaradi velike količine podatkov smo pripravili 1.000 vzorcev celotne podatkovne zbirke z izborom 1 % celotne podatkovne zbirke. Na predpripravljenih vzorcih podatkovne zbirke smo pognali funkcijo ReliefFexpRank v programskem paketu R. Izbor atributov je bil opravljen na računski gruči Arnesa, ki je del Slovenske iniciative za nacionalni grid[1].

Iz rezultatov smo tako izbrali 20 atributov, ki najbolje opišejo problemsko domeno in so predstavljeni na sliki 4.1.

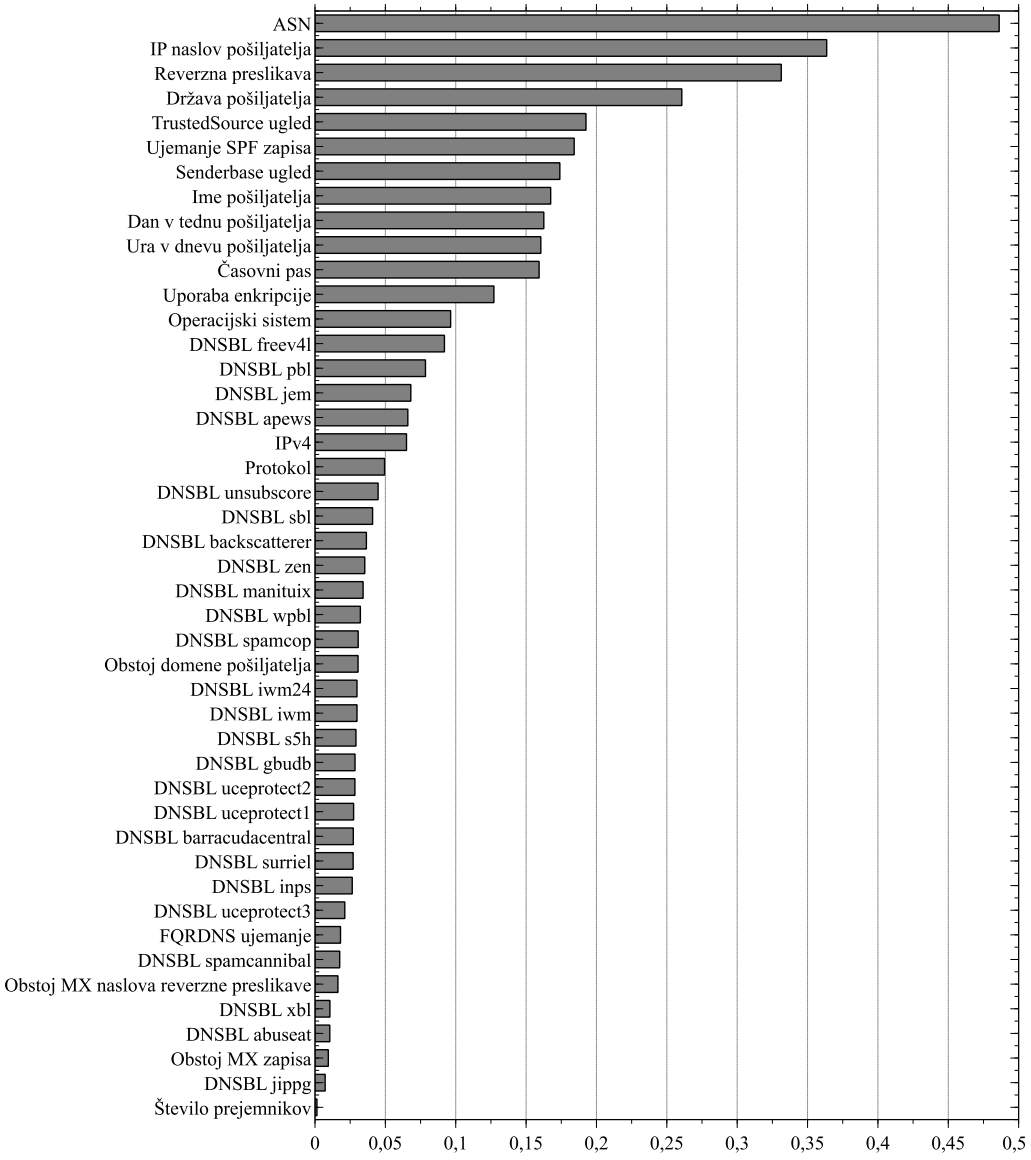
Ti atributi so:

- ASN
- DNSBL apews
- DNSBL freev4l
- DNSBL jem
- DNSBL pbl
- DNSBL unsubscore
- dan v tednu pošiljatelja
- država pošiljatelja
- naslov IP pošiljatelja
- IPv4
- ime pošiljatelja
- operacijski sistem
- protokol
- reverzna preslikava
- ugled Senderbase
- ugled TrustedSource
- ujemanje zapisa SPF
- uporaba enkripcije
- ura v dnevu pošiljatelja
- časovni pas

Izbor zgoraj predstavljenih atributov potrjuje naslednje, v strokovni literaturi predstavljene domneve:

[1]Slovenska iniciativa za nacionalni grid – SLING – `http://www.sling.si`

- atribut *ASN*, ki predstavlja ponudnika internetnih storitev, potrjuje, da obstaja močna povezanost med ponudniki internetnih storitev in pošiljatelji neželene elektronske pošte [52];
- vključitev atributov *DNSBL apews, DNSBL freev4l, DNSBL jem, DNSBL*



Slika 4.1: Porazdelitev atributov po pomembnosti glede na funkcijo ReliefFexpRank (prikazane so vrednosti ocene atributov)

pbl, DNSBL unsubscore, ugled TrustedSource, ugled Senderbase potrdi kakovost teh DNSBL podatkovnih zbirk in zbirk ugleda naslova IP;

- atributi *dan v tednu pošiljatelja, ura v dnevu pošiljatelja, časovni pas* potrjujejo domnevo, da večino neželene elektronske pošte pošljejo okuženi računalniki, ki so večinoma vključeni v omrežje v delovnem času [2];
- vključitev atributa *IPv4* razloži hipotezo, da se še vedno pošlje več neželene elektronske pošte preko IPv4 naslovnega prostora [53];
- *ime pošiljatelja* se v kampanjah, ki jih izvajajo pošiljatelji neželene elektronske pošte, ne spreminja [5];
- atribut *operacijski sistem* potrjuje domnevo o povezavi med neželeno elektronsko pošto in operacijskim sistemom pošiljatelja [54];
- vključitev atributov *protokol, uporaba enkripcije* in *ujemanje zapisa SPF* potrdi domnevo, da obstaja močna povezava med spoštovanjem protokola SMTP in neželeno elektronsko pošto. Pošiljatelji neželene elektronske pošte velikokrat zanemarijo spoštovanje protokolov v nameri, da bi v najkrajšem možnem času poslaličim večje število neželene elektronske pošte [55].

4.2 Izbor in primerjava algoritmov za učenje iz stacionarnih podatkov in algoritmov za učenje iz podatkovnih tokov

V problemski domeni označevanja neželene elektronske pošte želimo z uporabo različnih pristopov pravilno uvrstiti elektronska sporočila v dva razreda: želena in neželena elektronska pošta.

Med najbolj uspešne pristope štejemo uporabo algoritmov za učenje iz stacionarnih podatkov. V literaturi [6] se med največkrat uporabljanimi algoritmi za klasifikacijo elektronske pošte pojavljata dva algoritma: naivni Bayesov klasifikator [56] in odločitvena drevesa C4.5 [57].

Z uporabo učenja iz podatkovnih tokov želimo primerjati klasifikacijsko točnost z algoritmi za paketno učenje iz stacionarnih podatkov. Za klasifikacijske probleme pri metodah za učenje iz podatkovnih tokov najpogosteje uporabljamo dva algoritma: VFDT [46] in cVFDT [47]. S primerjavo klasifikacijskih točnosti zgoraj omenjenih algoritmov na naši podatkovni množici, sestavljeni iz atributov iz poglavja 4.1, želimo ugotoviti, kako uspešna sta algoritma za učenje iz podatkovnih tokov v primerjavi z izbranima algoritmoma za paketno učenje iz stacionarnih podatkov na primeru označevanja neželene elektronske pošte. Večina orodij, ki uporabljajo zgoraj omenjene algoritme za uvrščanje, uporablja vse atribute elektronske pošte, tudi same vsebine elektronske pošte, ki smo jo v našem pristopu namenoma izpustili.

Metode za paketno učenje iz stacionarnih podatkov smo uporabili, da smo izračunali ciljni razred na zbrani podatkovni množici, predstavljeni v poglavju 3. Pri tem smo poskušali posnemati realno okolje pri ponudniku elektronske pošte na način, da smo prilagodili velikost učne množice. Učna množica je bila tako sestavljena iz podatkov o elektronski pošti izpred določenega števila dni. S tem smo simulirali izdelavo odločitvenega modela, ki se lahko zgradi pri ponudniku elektronske pošte enkrat dnevno.

Ker je naša podatkovna zbirka ciljni razred že imela, smo le-tega uporabili za izračun klasifikacijske točnosti uporabljenega algoritma. Le to smo izračunali za vsak posamezni dan v podatkovni zbirki in s tem primerjali, kako se klasifikacijska točnost spreminja skozi čas. Na koncu smo izračunali povprečno klasifikacijsko točnost in standardni odklon. S tem smo pridobili povprečno dnevno klasifikacijsko točnost, ki smo jo lahko primerjali med različnimi uporabljenimi algoritmi in tudi z metodami inkrementalnega učenja iz podatkovnih tokov.

Pri uporabi metod inkrementalnega učenja iz podatkovnih tokov smo prevedli prejem elektronske pošte v podatkovni tok, kjer vsaka elektronska pošta predstavlja učni primer v podatkovnem toku.

Uporabili smo dve najbolj pogosti metodi za inkrementalno učenje iz podatkovnih tokov: uporaba algoritmov VFDT [46] in cVFDT [47]. Algoritem

VFDT inkrementalno gradi odločitveno drevo, obenem pa upošteva omejitve učenja iz podatkovnih tokov. Algoritem cVFDT je nadgradnja algoritma VFDT, ki bolje obravnava zvezne atribute in gradi vzporedno drevo VFDT algoritmu in takoj, ko obstoječe drevo ne dosega več dobre klasifikacijske točnosti, le-tega zamenja z vzporednim.

Algoritma smo poganjali z uporabo metode, kjer se za vsak primer v podatkovnem toku določi ciljni razred in takoj v naslednjem koraku algoritem ta podatek uporabi v svoji učni množici in s tem inkrementalno poveča odločitveni model.

Oba algoritma za inkrementalno učenje iz podatkovnih tokov smo uporabili brez kakršnihkoli omejitev učnega okna.

Pri uporabi algoritmov za inkrementalno učenje iz podatkovnih tokov VFDT in cVFDT smo zaradi računske zahtevnosti celotnega opravila klasifikacijsko točnost izračunali po klasifikaciji vsakih 10.000 primerov. V povprečju smo tako dobili več kot 20 povprečnih klasifikacijskih točnosti za vsak dan v podatkovnem toku.

Za izvajanje algoritmov paketnega učenja iz stacionarnih podatkov smo uporabili programski paket MOA[2], ki temelji na programski opremi Weka[3]. Za izvajanje obeh algoritmov za inkrementalno učenje iz podatkovnih tokov smo za osnovo izbrali programsko opremo streamDM-C++[4], ki implementira oba algoritma v programskem jeziku C in s tem pohitri samo izvajanje algoritmov.

4.2.1 Uporaba algoritma naivni Bayes

Algoritem naivnega Bayesa smo uporabili na način, da smo za učno množico uporabili bodisi podatke preteklega dne, podatke izpred sedmih dni, preteklega tedna bodisi preteklih 14 dni. Na tak način smo simulirali okolje, kjer

[2]MOA - Massive Online Analysis - `http://moa.cms.waikato.ac.nz/`

[3]Weka - `http://www.cs.waikato.ac.nz/ml/weka/`

[4]streamDM-C++: C++ Stream Data Mining: `https://github.com/huawei-noah/streamDM-Cpp`

se sistem uči samo iz podatkov iz preteklega časovnega okvirja.

Izbor časovnega okna za učno množico smo opravili ob predpostavki, da je to dovolj velika zbirka podatkov, ki jo lahko obdelamo v kratkem času in iz tega izdelamo dovolj dober odločitveni model, ki bo upošteval dovolj lastnosti neželene elektronske pošte.

Rezultati uporabe algoritma za paketno učenje iz stacionarnih podatkov naivni Bayes se nahajajo v tabeli 4.1. Iz rezultatov povprečne klasifikacijske točnosti lahko vidimo, da dosežemo najvišjo točnost v primeru, ko izberemo učno množico izpred zadnjih sedmih dni. Obenem je tudi standardni odklon najnižji in zato je izbira take množice najbolj primerna, ko gradimo odločitveni model z uporabo algoritma naivni Bayes.

Učna množica	Povprečna klasifikacijska točnost	Standardni odklon
podatkovna množica preteklega dne	93.2768%	2.5468
podatkovna množica enega dne izpred enega tedna	95,806%	1,7779
podatkovna množica izpred zadnjih sedmih dni dne	96,1356%	1,4910
podatkovna množica izpred zadnjih štirinajstih dni dne	96,0869%	1,5508

Tabela 4.1: Povprečna klasifikacijska točnost in standardni odklon algoritma naivni Bayes pri uporabi različnih učnih množic

4.2.2 Uporaba algoritma C4.5

Podobno kot algoritem naivnega Bayesa smo uporabili tudi algoritem C4.5. Zaradi omejenih računskih zmogljivosti smo algoritem C4.5 omejili na učno množico izpred enega in sedmih dni. Uporaba večjih učnih množic je namreč presegla spominske zmogljivosti računalnikov, ki smo jih lahko uporabili.

Razlog za tak izbor učnih množic izhaja iz rezultatov uporabe algoritma naivni Bayes, kjer smo z učno množico izpred enega dne dosegli nižjo klasifikacijsko točnost kot z učno množico izpred enega tedna. Količina in vsebina elektronske pošte se namreč precej ujema z običajnim delovnim časom, kar pomeni, da je učna množica izpred enega dne manj reprezentativna (delovnik, dela prost dan) kot učna množica izpred sedmih dni.

Rezultati uporabi algoritma za paketno učenje iz stacionarnih podatkov C4.5 se nahajajo v tabeli 4.2. Z uporabo učne množice izpred enega dne dobimo najvišjo klasifikacijsko točnost algoritma C4.5, ki tudi presega klasifikacijsko točnost algoritma naivni Bayes, ki je predstavljena v tabeli 4.1.

Učna množica	Povprečna klasifikacijska točnost	Standardni odklon
podatkovna množica preteklega dne	95.1135%	2.0047
podatkovna množica enega dne izpred enega tedna	96.1470%	1.8947

Tabela 4.2: Povprečna klasifikacijska točnost in standardni odklon algoritma C4.5 pri uporabi različnih učnih množic

4.2.3 Uporaba algoritmov VFDT in cVFDT

Pri uporabi metod inkrementalnega učenja iz podatkovnih tokov smo uporabili algoritma VFDT in cVFDT brez omejitev velikosti učnega okna kot tudi brez omejitve zasedenega pomnilnika.

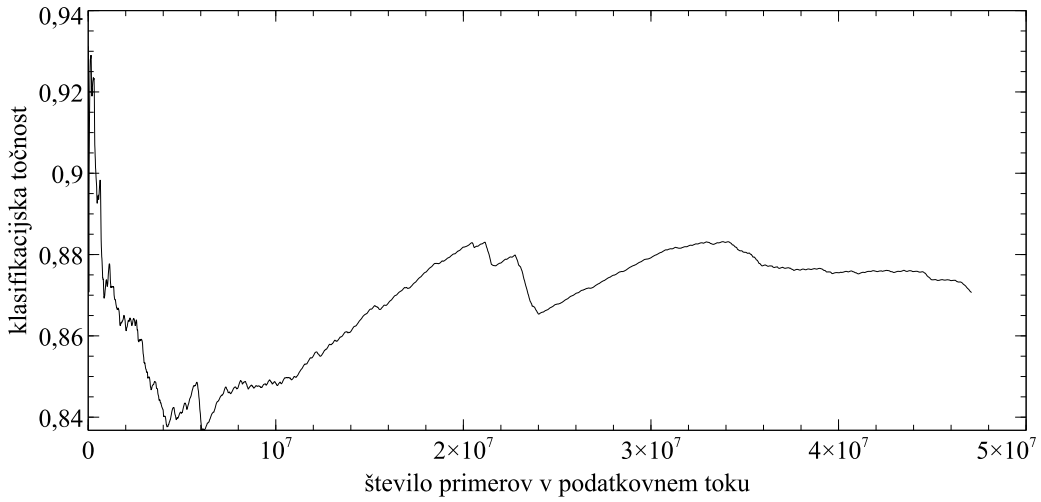
Povprečna klasifikacijska točnost algoritmov za inkrementalno učenje iz podatkovnih tokov je predstavljena v tabeli 4.3.

Povprečna klasifikacijska točnost se je pri obdelavi podatkovnega toka spreminjala, kar je lepo razvidno na grafu 4.2 pri uporabi algoritma VFDT in na grafu 4.3 pri uporabi algoritma cVFDT. V primeru uporabe algoritma cVFDT je iz grafa lepo vidno, da se je klasifikacijska točnost s povečevanjem števila učnih modelov dvigala in dosegla višjo povprečno vrednost kot

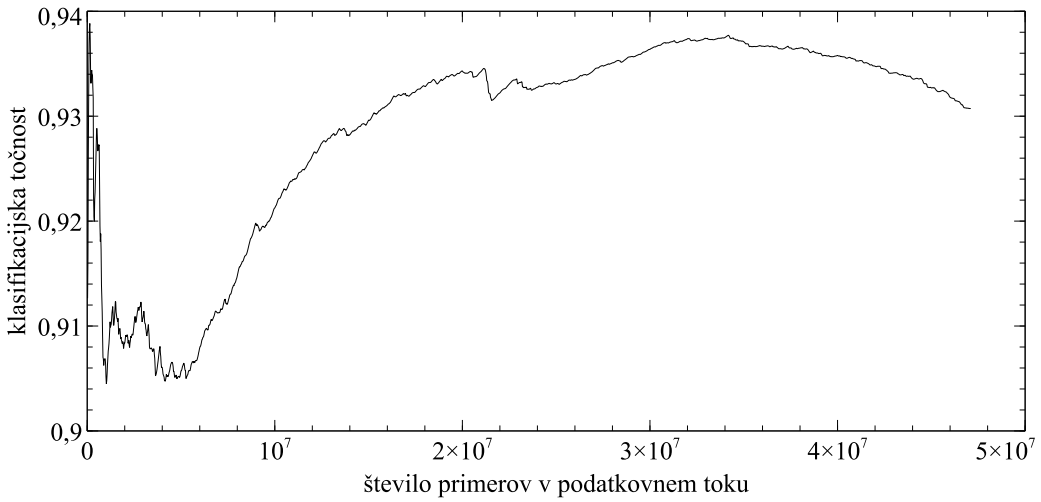
Algoritem	Povprečna klasifikacijska točnost	Standardni odklon
VFDT	86.9293%	1.3485
cVFDT	92.8897%	0.9687

Tabela 4.3: Povprečna klasifikacijska točnost in standardni odklon algoritmov VFDT in cVFDT brez omejitev

v primeru uporabe algoritma VFDT.



Slika 4.2: Klasifikacijska točnost algoritma VFDT brez omejitev



Slika 4.3: Klasifikacijska točnost algoritma cVFDT brez omejitev

4.2.4 Primerjava uporabljenih algoritmov

Pri analizi smo izbrali 4 algoritme: dva algoritma za paketno učenje iz stacionarnih podatkov in dva algoritma za inkrementalno učenje iz podatkovnih tokov.

Algoritma naivni Bayes in C4.5 sta na množici 20 najbolj reprezentativnih atributov pokazala povprečno klasifikacijsko točnost nad 93% , zbranih na način brez poseganja v samo vsebino elektronske pošte. Algoritem C4.5 je imel povprečno klasifikacijsko točnost celo nad 95%. Z izborom prave učne množice smo izboljšali klasifikacijsko točnost obeh algoritmov nad 96%. Taka klasifikacijska točnost je nižja od povprečne klasifikacijske točnosti sistema, s katerim smo zbrali celotno podatkovno zbirko in tudi od povprečne klasifikacijske točnosti pri uporabi metod blokiranja neželene elektronske pošte z uporabo spletnih podatkovnih baz IP ugleda [58]. Obenem je ta rezultat tudi precej slabši v primerjavi s komercialnimi produkti, ki večinoma dosegajo klasifikacijsko točnost nad 99% [59].

Algoritem za inkrementalno učenje iz podatkovnih tokov VFDT je dosegel klasifikacijsko točnost 86.9%, medtem ko je algoritem cVFDT dosegel klasikacijsko točnost 92.8%.

V primerjavi z algoritmoma za paketno učenje iz stacionarnih podatkov se izkažeta algoritma za inkrementalno učenje iz podatkovnih tokov za slabša in v okolju ponudnika elektronske pošte precej neuporabna. S smiselno izbiro učne množice lahko algoritma naivni Bayes in C4.5 prilagodimo, da imata boljšo klasifikacijsko točnost od algoritmov za inkrementalno učenje iz podatkovnih tokov ob pravilni izbiri učne množice.

4.3 Izbor velikosti učnega okna pri učenju iz podatkovnih tokov

Pri algoritmih inkrementalnega učenja iz podatkovnih tokov je zelo pomemben dejavnik tudi velikost učnega okna. Ta pogojuje klasifikacijsko točnost samega algoritma kot tudi računsko in pomnilniško obremenitev računalnika.

Z namenom optimizacije strojnih resursov računalnika smo poskušali identificirati optimalno velikost učnega okna za oba algoritma, ki smo ju izbrali v poglavju 4.2. S tem smo omogočili manjšo zasedenost pomnilnika in hitrejše izvajanje algoritma na primeru problemske domene označevanja neželene elektronske pošte.

Za identifikacijo optimalnega učnega okna smo uporabili isto programsko opremo in podatkovno zbirko kot v poglavju 4.2. Primerjali smo povprečno klasifikacijsko točnost algoritmov, ki smo jo izračunali v korakih po 10.000 primerkov iz podatkovne zbirke.

Glede na zahtevnost uporabe algoritmov VFDT in cVFDT na podatkovni zbirki smo izbrali tri velikosti učnega okna: 1.000, 10.000 in 100.000 primerov iz podatkovne zbirke. Vsaka od teh velikosti vsebuje določene elemente, ki bi nam lahko pomagali pri identifikaciji optimalnega učnega okna. Najmanjše učno okno nam tako lahko pove, ali se klasifikacijska točnost bistveno spremeni, če omejimo učno okno na samo 1.000 primerov. Z omejitvijo velikosti učnega okna na 100.000 primerov smo poskušali ugotoviti, ali s povečevanjem učnega okna izboljšamo povprečno klasifikacijsko točnost. Uporabili smo tudi vmesno velikost učnega okna 10.000 primerov.

Oba algoritma sta pri različnih velikostih učnega okna dala rezultate, ki so predstavljeni v tabeli 4.4.

Algoritem	Velikost učnega okna	Povprečna klasifikacijska točnost	Standardni odklon
VFDT	1.000	86.9226%	0.1095
VFDT	10.000	86.8330%	0.1016
VFDT	100.000	86.8876%	0.8334
cVFDT	1.000	93.4710%	0.0630
cVFDT	10.000	93.4540%	0.0490
cVFDT	100.000	93.45400%	0.0351

Tabela 4.4: Povprečna klasifikacijska točnost in standardni odklon algoritmov VFDT in cVFDT pri uporabi različnih velikosti učnega okna

Rezultati izbire optimalne velikosti učnega okna so pokazali, da velikost učnega okna ne vpliva bistveno na povprečno klasifikacijsko točnost algoritmov za inkrementalno učenje iz podatkovnih tokov. Že okno velikosti 1.000 primerov namreč doseže enako klasifikacijsko točnost kot vsako večje testirano okno. Razlike se pojavijo samo pri večjih razlikah v klasifikacijski točnosti pri procesiranju celotnega podatkovnega toka. Večje kot je učno okno, manj se klasifikacijska točnost spreminja skozi podatkovni tok.

Majhen vpliv velikosti učnega okna je precej nepričakovan, saj smo pričakovali, da bo učno okno velikosti 1.000 primerov imelo precej manjšo povprečno klasifikacijsko točnost kot večja okna. Zanimiva je tudi ugotovitev, da oba algoritma dosegata boljše klasifikacijske točnosti pri uporabi omejitve učnega okna kot pri uporabi algoritmov brez omejitev (tabela 4.3).

4.4 Primerjava z obstoječimi metodami

Uporaba metod strojnega učenja na podatkovni zbirki brez osebnih podatkov, ki je zbrana v transportni fazi prenosa elektronske pošte, je pristop, ki pri problemski domeni označevanja neželene elektronske pošte še ni bil upora-

bljen. Obenem uporaba algoritmov za inkrementalno učenje iz podatkovnih tokov predstavlja nov pristop k reševanju tega problema.

Vseeno lahko tak pristop primerjamo s preostalimi metodami zaznave neželene elektronske pošte s primerjanjem klasifikacijske točnosti.

Za primerjavo smo izbrali dve metodi, ki ju najdemo v literaturi in ju lahko primerjamo s pridobljenimi rezultati. Prva metoda je uporaba podatkovne zbirke DNSBL Zen projekta Spamhaus[5]. Podatkovna zbirka DNSBL Zen vsebuje naslove IP pošiljateljev neželene elektronske pošte in velja za najbolj kakovostno zbirko [60]. Uporaba te metode ne posega v osebne podatke prejemnika elektronske pošte in zato je primerna za uporabo pri ponudnikih storitev elektronske pošte.

Druga metoda, ki smo jo primerjali s pridobljenimi rezultati iz poglavij 4.2 in 4.4, je metoda uporabe algoritmov za inkrementalno učenje iz podatkovnih tokov GNUsmail [49], ki uporablja za učno množico tudi vsebino elektronske pošte.

Obe metodi smo primerjali s primerjavo klasifikacijskih točnosti med izbrano metodo in najboljšo klasifikacijsko točnostjo algoritmov naivni Bayes, C4.5, VFDT in cVFDT.

4.4.1 Primerjava z DNSBL podatkovno zbirko Zen

Podatkovna zbirka DNSBL Zen ima v literaturi [18, 60] najboljšo klasifikacijsko točnost med vsemi podatkovnimi zbirkami DNSBL. Ker smo podatke iz podatkovne zbirke DNSBL Zen že imeli v učni množici kot enega izmed atributov, smo lahko preverili ujemanje atributa s ciljnim razredom.

Iz rezultatov, predstavljenih v tabeli 4.5, lahko ugotovimo, da uporaba algoritmov za inkrementalno učenje iz podatkovnih tokov, kjer podatkovni tok vsebuje samo 20 atributov, omogoča boljšo klasifikacijsko točnost kot uporabo podatkovne zbirke DNSBL Zen. Uporaba podatkovne zbirke DNSBL Zen je vendarle dovolj natančna, da jo lahko uporabimo v začetni fazi prenosa elektronskega sporočila (vzpostavitev povezave, greylisting itd.), ker predsta-

[5]Spamhaus Project - `http://www.spamhaus.org`

Uporabljena metoda	Povprečna klasifikacijska točnost
algoritem naivni Bayes	96,0869%
algoritem C4.5	96.1470%
algoritem VFDT	86.9293%
algoritem cVFDT	93.4710%
podatkovna zbirka DNSBL Zen	89.2246%

Tabela 4.5: Povprečna klasifikacijska točnost algoritmov naivni Bayes, C4.5, VFDT, cVFDT in ujemanje atributa iz podatkovne zbirke DNSBL Zen s ciljnim razredom

vlja najmanjšo obremenitev resursov poštnega strežnika, medtem ko uporaba algoritmov za strojno učenje uporablja bolj resurse poštnega strežnika.

4.4.2 Primerjava s programsko opremo GNUsmail

Programska oprema GNUsmail[6] uporablja za inkrementalno učenje iz podatkovnih tokov več različnih algoritmov in za svojo učno množico uporablja celotno elektronsko sporočilo, vključno z osebnimi podatki, ki jih v naši podatkovni zbirki nismo uporabili.

Za primerjavo smo tako uporabili klasifikacijsko točnost, ki jo je programska oprema GNUsmail dosegla v članku "GNUSmail: Open framework for on-line email classification". V članku so Carmona-Cejudo in sod. [49] uporabili tri različne algoritme za inkrementalno učenje iz podatkovnih tokov in pri tem dosegli različne klasifikacijske točnosti za različne podatkovne tokove elektronskih sporočil. Za podatkovni tok so izbrali zbirko elektronski sporočil Enron [61].

V članku so tako uporabili naslednje algoritme za leno učenje in VFDT. Med njimi se je najbolj izkazal algoritem najbližjih sosedov z uporabo nepovezanih splošnih primerov (angl. *Nearest-neighbor-like algorithm using non-nested generalized exemplars*), ki je dosegel povprečno klasifikacijsko točnost 78.6795%. Povprečna klasifikacijska točnost algoritmov, ki smo jih uporabili

[6]GNUsmail - `https://github.com/mbaena/gnusmail`

v poglavju 4.2, je tako presegla povprečno klasifikacijsko točnost v članku uporabljenega algoritma.

Razlog za to se skriva v količini elektronske pošte, ki jo algoritmi potrebujejo, da dosežejo visoko klasifikacijsko točnost. Algoritma VFDT in cVFDT se namreč najbolje obneseta v primerih, ko podatkovni tok vsebuje veliko število primerov, kar v primeru naše podatkovne množice zagotovo je. V članku so namreč uporabili podatkovno množico z 2479 elektronskimi sporočili, kar se je odražalo v nižji klasifikacijski točnosti uporabljenega algoritma.

Poglavje 5

Zaključek

V magistrski nalogi smo predstavili napreden način zaznave neželene elektronske pošte z uporabo algoritmov za inkrementalno učenje iz podatkovnih tokov. Pri tem smo upoštevali omejitve zaščite osebnih podatkov in pri gradnji odločitvenega modela uporabljali samo atribute, ki jih lahko pridobimo v transportni fazi prenosa elektronskega sporočila.

Z upoštevanjem zaščite osebnih podatkov smo sestavili podatkovno zbirko 47.073.461 zapisov elektronskih sporočil, ki so bili opisani s 60 atributi. Izmed 60 atributov smo izbrali 20 takih, ki najbolj vplivajo na odločitev, ali elektronsko sporočilo uvrstimo med želena ali neželena sporočila.

V nadaljevanju smo primerjali metodi paketnega učenja iz stacionarnih podatkov z metodami inkrementalnega učenja iz podatkovnih tokov in ugotovili, da metode paketnega učenja iz stacionarnih podatkov ob izbiri optimalne učne množice doseže boljšo povprečno klasifikacijsko točnost kot metode za inkrementalno učenje iz podatkovnih tokov. Obenem smo ugotovili, da algoritmi za paketno učenje iz stacionarnih podatkov dosegajo klasifikacijsko točnost nad 90%, če v fazi učenja uporabimo učno množico izpred zadnjih nekaj dni. Na drugi strani tudi algoritmi za inkrementalno učenje iz podatkovnih tokov dosegajo klasifikacijsko točnost nad 90% in pri tem potrebujejo manjše število primerov v učnem oknu ter porabijo manj strojnih resursov.

Iz opravljenih eksperimentov ni vidno, da bi bila metoda inkrementalnega

učenja iz podatkovnih tokov boljša od preostalih, danes uporabljenih metod za uvrščanje elektronske pošte.

Z nadaljnim raziskovanjem optimalne velikosti učnega okna, razvojem novih algoritmov za inkrementalno učenje iz podatkovnih tokov in boljšim poznavanjem problemske domene bomo lahko približali učinkovitost zaznave neželene pošte trenutno aktualnim metodam in obenem ohranili zasebnost prejemnika in pošiljatelja.

Poglavje 6

Dodatki

6.1 Tabela uporabljenih atributov

Tabela 6.1: Tabela uporabljenih atributov

ime atributa	oblika atributa	opis atributa
asn	diskretni	Podatek o ponudniku interneta glede na pošiljateljev naslov IP, pridobljen iz podatkovne baze MaxMind GeoLite asn[1]
Časovni pas	diskretni	Podatek o časovnem pasu naslova IP, pridobljen iz podatkovne baze Maxmind GeoIP City2[2]
Dan v tednu	zvezni	Podatek o dnevu v tednu prejema elektronskega sporočila.

[1] `http://dev.maxmind.com/geoip/legacy/geolite/`
[2] `https://www.maxmind.com/en/geoip2-city`

Dan v tednu pošiljatelja	zvezni	Lokalni dan v tednu pošiljatelja, pridobljen iz geolokacije pošiljateljevega naslova IP in časa prejema elektronskega sporočila
DNSBL abuseat	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: cbl.abuseat.org
DNSBL apews	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: l2.apews.org
DNSBL backscatterer	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: ips.backscatterer.org
DNSBL barracudacentral	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: b.barracudacentral.org
DNSBL fabel	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: spamsources.fabel.dk
DNSBL freev4l	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: free.v4bl.org
DNSBL gbudb	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: truncate.gbudb.net

DNSBL inps	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl.inps.de
DNSBL iwm	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: sip.invaluement.com
DNSBL iwm24	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: sip24.invaluement.com
DNSBL jem	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: hostkarma.junkemailfilter.com
DNSBL jippg	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: mailabuse.blacklist.jippg.org
DNSBL manituix	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: ix.dnsbl.manitu.net
DNSBL pbl	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: pbl.spamhaus.org
DNSBL s5h	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: all.s5h.net

DNSBL sbl	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: sbl.spamhaus.org
DNSBL sem	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: bl.spameatingmonkey.net
DNSBL senderscore	zvezni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: score.senderscore.com
DNSBL services	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: korea.services.net
DNSBL sorbs	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl.sorbs.net
DNSBL spamcannibal	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: bl.spamcannibal.org
DNSBL spamcop	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: bl.spamcop.net
DNSBL spamrats	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: all.spamrats.com

DNSBL spamratsdyna	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dyna.spamrats.com
DNSBL spamratsspam	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: spam.spamrats.com
DNSBL surriel	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: psbl.surriel.com
DNSBL swinog	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsrbl.swinog.ch
DNSBL uceprotect0	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl-0.uceprotect.net
DNSBL uceprotect1	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl-1.uceprotect.net
DNSBL uceprotect2	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl-2.uceprotect.net
DNSBL uceprotect3	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: dnsbl-3.uceprotect.net

DNSBL unsubscore	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: ubl.unsubscore.com
DNSBL wpbl	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: db.wpbl.info
DNSBL xbl	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: xbl.spamhaus.org
DNSBL zen	binarni	Podatek o prisotnosti pošiljateljevega naslova IP v podatkovni zbirki DNSBL: zen.spamhaus.org
Država pošiljatelja	diskretni	Podatek o državi, iz katere prihaja naslov IP pošiljatelja, pridobljen iz podatkovne baze MaxMind GeoLite Country[3]
FQRDNS ujemanje	binarni	Podatek o tem ali se obratna preslikava DNS pošiljateljevega naslova IP ujema s tistimi, ki bi za pošiljanje morali uporabiti SMTP strežnike njihovih ponudnikov interneta.[4]
SMTP HELO ime	diskretni	Ime pošiljatelja iz poizvedbe SMTP HELO

[3] `http://dev.maxmind.com/geoip/legacy/geolite/`

[4] `https://github.com/stevejenkins/hardwarefreak.com-fqrdns.pcre/blob/master/fqrdns.pcre`

Ime pošiljatelja	diskretni	Ime pošiljatelja iz ovojnice elektronskega sporočila
Pošiljateljev Naslov IPv4	diskretni	Pošiljateljev naslov IP
IPv4	binarni	Podatek o tem ali pošiljatelj uporablja IPv4 naslov ali IPv6 naslov
Lokalni čas pošiljatelja	zvezni	Lokalni čas pošiljatelja, pridobljen iz geolokacije pošiljateljevega naslova IP in časa prejema elektronskega sporočila
Obstoj domene pošiljatelja	binarni	Podatek o tem, ali domena pošiljatelja, pridobljena v ovojnici, obstaja
Obstoj MX naslova obratne preslikave	binarni	Podatek o tem, ali ima obratna preslikava DNS pošiljateljevega naslova IP veljaven zapis MX
Obstoj MX zapisa	binarni	Podatek o tem, ali ima pošiljatelj veljaven zapis MX
Operacijski sistem	diskretni	Podatek o operacijskem sistemu pošiljatelja, pridobljen preko sistema p0f. Več na strani 3
Protokol	diskretni	Podatek o uporabljenem protokolu SMTP pri prenosu sporočila: ESMTP ali SMTP
Obratna preslikava DNS	binarni	Podatek o tem ali ima pošiljateljevega naslov IP obratno preslikavo DNS

ugled Senderbase	zvezni	Podatek o ugledu pošiljateljevega naslova IP v podatkovni zbirki Senderbase[5]
Število prejemnikov	zvezni	Podatek o številu prejemnikov elektronskega sporočila
ugled TrustedSource	zvezni	Podatek o ugledu pošiljateljevega naslova IP v podatkovni bazi TrustedSource[6]
Ujemanje imena pošiljatelja in naslova IP	binarni	Podatek o tem, ali se obratna preslikava DNS pošiljateljevega naslova IP, v katerega se preslika ime pošiljatelja, ujema z imenom pošisljatelja
Ujemanje obratne preslikave in naslova IP	binarni	Podatek o tem, ali se obratna preslikava DNS pošiljateljevega naslova IP ujema s pošiljateljevem naslovom IP
Ujemanje SPF zapisa	binarni	Podatek o ujemanju SPF [35] zapisa
Uporaba enkripcije enkripcija	binarni	Podatek ali je bila pri prenosu sporočila uporabljena
Ura	zvezni	Podatek o uri prejema elektronskega sporočila
Ura pri pošiljatelju	zvezni	Podatek o uri pošiljanju pri pošiljatelju, pridobljen iz geolokacije pošiljateljevega naslova IP in časa prejema elektronske sporočila

[5] `http://www.senderbase.org`[62]
[6] `http://trustedsource.org/`

Literatura

[1] Inc. The Radicati Group. *Email Statistics Report, 2015-2019*. Teh. poročilo. 2015, str. 4.

[2] Symantec. *Symantec Internet Security Threat Report*. Teh. poročilo April. 2015, str. 119.

[3] J Klensin. *Simple Mail Transfer Protocol*. RFC 5321 (Draft Standard). Okt. 2008. URL: http://www.ietf.org/rfc/rfc5321.txt.

[4] Justin M Rao in David H Reiley. "The Economics of Spam". V: *Journal of Economic Perspectives* 26.3 (2012), str. 87–110. ISSN: 0895-3309. DOI: 10.1257/jep.26.3.87.

[5] Sérgio S.C. Silva in sod. "Botnets: A survey". V: *Computer Networks* 57.2 (feb. 2013), str. 378–403. ISSN: 13891286. DOI: 10.1016/j.comnet.2012.07.021.

[6] Gordon V. Cormack. "Email Spam Filtering: A Systematic Review". V: *Foundations and Trends in Information Retrieval* 1.4 (2008), str. 335–455. ISSN: 1554-0669. DOI: 10.1561/1500000006.

[7] Jernej Porenta in Mojca Ciglarič. "Comparing commercial IP reputation databases to open-source IP reputation algorithms". V: *International Journal of Computer Systems Science and Engineering* 28.1 (2013), str. 53–66.

[8] Holly Esquivel, Aditya Akella in Tatsuya Mori. "On the effectiveness of IP reputation for spam filtering". V: *2010 Second International Conference on COMmunication Systems and NETworks (COMSNETS 2010)* (jan. 2010), str. 1–10. DOI: 10.1109/COMSNETS.2010.5431981.

[9] A Borg in N Lavesson. "E-mail Classification using Social Network Information". V: *Seventh International Conference on Availability, Reliability and Security (ARES)* (2012), str. 168–173.

[10] Bradley Taylor. "Sender reputation in a large webmail service". V: *Proceedings of the Third Conference on Email and Anti-Spam (CEAS)*. 2006, str. 1–6.

[11] Barry Leiba in Jim Fenton. "DomainKeys Identified Mail (DKIM): Using digital signatures for domain verification". V: *Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS)*. 2007, str. 6–11.

[12] E Blanzieri in A Bryl. "A Survey of Learning-Based Techniques of Email Spam Filtering". V: *Artificial Intelligence Review* 29.1 (2008), str. 63–92.

[13] JM Carmona in M Baena-García. "Using GNUsmail to Compare Data Stream Mining Methods for On-line Email Classification." V: *JMLR: Workshop and Conference Proceedings–Workshop on Applications of Pattern Analysis*. Zv. 17. 2011, str. 12–18.

[14] MPS Bogawar in KK Bhoyar. "Email Mining: A Review". V: *International Journal of Computer Science Issues* 9.1 (2012), str. 429–434.

[15] Alexandre Bronstein, Joydip Das in Marsha Duro. "Self-Aware Services : Using Bayesian Networks for Detecting Anomalies in internet-based Services". V: *2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings*. 2001, str. 623–638.

[16] CE Drake, JJ Oliver in EJ Koontz. "Anatomy of a phishing email". V: *Conference on Email and Anti-Spam*. 2004, str. 1–8.

[17] Meizhen Wang in sod. "Research on Behavior Statistic Based Spam Filter". V: *2009 First International Workshop on Education Technology and Computer Science*. Ieee, 2009, str. 687–691. ISBN: 978-0-7695-3557-9. DOI: 10.1109/ETCS.2009.413.

[18] Sushant Sinha, Michael Bailey in Farnam Jahanian. "Improving spam blacklisting through dynamic thresholding and speculative aggregation". V: *Proc. of 17th NDSS*. 2010, str. 104–119.

[19] Yuhong Liu in Y Sun. "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis". V: *Proc. of 2nd IEEE Int. Conference on Social Computing*. 2010, str. 65–72.

[20] Christian Kreibich, C Kanich in K Levchenko. "Spamcraft: An inside look at spam campaign orchestration". V: *Proc. of 2nd USENIX LEET*. September. 2009, str. 4–8.

[21] HY Lam in DY Yeung. "A learning approach to spam detection based on social networks". V: *Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS)*. 2007, str. 18–24.

[22] Ronald Nussbaum, Abdol-Hossein Esfahanian in Pang-Ning Tan. "History-Based Email Prioritization". V: *2009 International Conference on Advances in Social Network Analysis and Mining*. Ieee, jul. 2009, str. 364–365. ISBN: 978-0-7695-3689-7. DOI: 10.1109/ASONAM.2009.44.

[23] *SpamAssassin*. 2016. URL: http://spamassassin.apache.org/ (pridobljeno 10. 2. 2016).

[24] B Leiba in sod. "SMTP Path Analysis." V: *CEAS* 2.1 (2005), str. 54–66.

[25] *Mail abuse prevention system - MAPS*. 1996. URL: http://www.mail-abuse.com/ (pridobljeno 22. 10. 2015).

[26] *Blacklists compared*. URL: https://www.sdsc.edu/%7B~%7Djeff/spam/Blacklists%7B%5C_%7DCompared.html (pridobljeno 7. 1. 2016).

[27] Nathan Dimmock in Ian Maddison. "Peer-to-peer collaborative spam detection". V: *Crossroads* 11.2 (dec. 2004), str. 4–8. ISSN: 15284972. DOI: 10.1145/1144403.1144407.

[28] Ernesto Damiani in DC di Vimercati. "A reputation-based approach for choosing reliable resources in peer-to-peer networks". V: *Proceedings of the 9th ACM conference on Computer and communications security*. 2002, str. 207–216. ISBN: 1581136129.

[29] A Kołcz. "The impact of feature selection on signature-driven spam detection". V: *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS-2004)* (2004), str. 18–23.

[30] A Kołcz in A Chowdhury. "Hardening fingerprinting by context". V: *Proceedings of the fourth international conference on email and anti-spam*. 3.1 (2007), str. 48–62.

[31] Mike Burrows Frank Dabek Ted Wobber Martin Abadi Andrew Birrell. "Bankable Postage for Network Services". V: *Proceedings of the 8th Asian Computing Science Conference*. Mumbai, India: Springer-Verlag, dec. 2003, str. 72–90.

[32] J Blosser in D Josephsen. "Scalable centralized bayesian spam mitigation with bogofilter". V: *Proceedings of the 18th USENIX conference on System administration* 1.18 (2004), str. 1–20.

[33] C Dwork in M Naor. "Pricing via processing or combatting junk mail". V: *Advances in Cryptology—CRYPTO'92* 740 (1992), str. 139–147.

[34] JR Levine. "Experiences with greylisting". V: *In Second Conference on Email and Anti-Spam* 2 (2005), str. 68–74.

[35] M Wong in W Schlitt. "Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1". V: (2006). URL: http://www.hjp.at/doc/rfc/rfc4408.html.

[36] J Lyon in M Wong. *Sender ID: Authenticating E-Mail*. Teh. poročilo 4406. 2006. URL: http://www.ietf.org/rfc/rfc4406.txt.

[37] M Delany. "Domain-based email authentication using public keys advertised in the DNS (DomainKeys)". V: (2007). URL: `http://tools.ietf.org/html/rfc4870.txt`.

[38] M Kucherawy in E Zwicky. "Domain-based message authentication, reporting, and conformance (DMARC)". V: (2015). URL: `http://tools.ietf.org/html/rfc7489.txt`.

[39] B Agrawal, N Kumar in M Molle. "Controlling spam emails at the routers". V: *IEEE International Conference on Communications* 3 (2005), str. 1588–1592.

[40] CC Lai in MC Tsai. "An empirical performance comparison of machine learning methods for spam e-mail categorization". V: *HIS'04. Fourth International Conference on Hybrid Intelligent Systems* (2004), str. 44–48.

[41] T Fawcett. "In vivo spam filtering: a challenge problem for KDD". V: *ACM SIGKDD Explorations Newsletter* 5.2 (2003), str. 140–148.

[42] J Chan, I Koprinska in J Poon. "Co-Training on Textual Documents with a Single Natural Feature Set." V: *ADCS* (2004), str. 47–54.

[43] JMM da Cruz in GV Cormack. "Using old spam and ham samples to train email filters". V: *Sixth Conference on Email and Anti-Spam - CEAS*. 2009, str. 32–44.

[44] Gordon V Cormack in Jose-Marcio da Cruz. "On the relative age of spam and ham training samples for email filtering". V: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, str. 744–745.

[45] I Androutsopoulos, EF Magirou in DK Vassilakis. "A Game Theoretic Model of Spam E-Mailing." V: *CEAS* 2.1 (2005), str. 48–56.

[46] Pedro Domingos in Geoff Hulten. "Mining high-speed data streams". V: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00* (2000), str. 71–80. DOI: 10.1145/347090.347107.

[47] Geoff Hulten, Laurie Spencer in Pedro Domingos. "Mining time-changing data streams". V: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01* (2001), str. 97–106. DOI: 10.1145/502512.502529.

[48] Haixun Wang in sod. "Mining concept-drifting data streams using ensemble classifiers". V: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. New York, New York, USA: ACM Press, avg. 2003, str. 226–246. ISBN: 1581137370. DOI: 10.1145/956750.956778.

[49] José M. Carmona-Cejudo in sod. "GNUSmail: Open framework for on-line email classification". V: *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. 2011, str. 1141–1142. ISBN: 9781607506065. DOI: 10.3233/978-1-60750-606-5-1141.

[50] Marko Robnik-Šikonja in Igor Kononenko. "Theoretical and Empirical Analysis of ReliefF and RReliefF". V: *Machine Learning* 53.1-2 (2003), str. 23–69. ISSN: 08856125. DOI: 10.1023/A:1025667309714. arXiv: 0005074v1 [arXiv:astro-ph].

[51] M Robnik-Šikonja in P Savicky. *CORElearn—Classification, Regression, Feature Evaluation and Ordinal Evaluation*. 2012. URL: http://lkm.fri.uni-lj.si/rmarko/software/ (pridobljeno 23. 5. 2016).

[52] Giovane C M Moura in sod. "Internet bad neighborhoods aggregation". V: *Proceedings of the 2012 IEEE Network Operations and Management Symposium, NOMS 2012*. 2012, str. 343–350. ISBN: 9781467302685. DOI: 10.1109/NOMS.2012.6211917.

[53] Jakub Czyz in sod. "Don't Forget to Lock the Back Door! A Characterization of IPv6 Network Security Policy". V: *Network and Distributed System Security Symposium (2016)* February (2016), str. 21–24. DOI: 10.14722/ndss.2016.23047.

[54] Georgios Kakavelakis, Robert Beverly in Joel Young. "Auto-learning of SMTP TCP Transport-Layer Features for Spam and Abusive Message Detection". V: *Proceedings of the 25th USENIX Large Installation Systems Administration Conference (LISA)*. Dec. 2011, str. 164–178.

[55] Pin-Ren Chiou, Po-Ching Lin in Chun-Ta Li. "Blocking spam sessions with greylisting and block listing based on client behavior". English. V: (), str. 184–189. ISSN: 1738-9445.

[56] Gerd Brewka. *Artificial intelligence—a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. Series in Artificial Intelligence, Englewood Cliffs, NJ.* Zv. 11. 01. 1996, str. 78. ISBN: 0136042597. DOI: 10.1017/S0269888900007724. arXiv: 9809069v1 [arXiv:gr-qc].

[57] J Ross Quinlan. *C4.5: Programs for Machine Learning*. Zv. 1. 3. 1992, str. 302. ISBN: 1558602380. DOI: 10.1016/S0019-9958(62)90649-6.

[58] Jernej Porenta in Mojca Ciglarič. "Empirical comparison of IP reputation databases". V: *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*. CEAS '11. New York, NY, USA: ACM, 2011, str. 220–226. ISBN: 978-1-4503-0788-8.

[59] Martijn Grooten in Ionuţ Răileanu. *VBSpam comparative review*. Teh. poročilo March. Virus Bulletin, 2016. URL: https://www.virusbulletin.com/uploads/pdf/magazine/2016/201603-vbspam-comparative.pdf.

[60] *Comparison of other DNS blacklists and DNS whitelists / Statistics*. URL: http://dnsbl.inps.de/index.cgi (pridobljeno 23. 8. 2016).

[61] Bryan Klimt in Yiming Yang. "The Enron Corpus: A New Dataset for Email Classification Research". V: Springer Berlin Heidelberg, 2004, str. 217–226. DOI: 10.1007/978-3-540-30115-8_22.

[62] *SenderBase Reputation Score*. Teh. poročilo. Cisco Ironport, 2006. URL: http://www.cisco.com/c/en/us/support/docs/security/email-security-appliance/118380-technote-esa-00.pdf.